Yelp Project:

Optimal Restaurant Location Identification using Yelp and other data sets

Linxi Wu
Jiayi Du
Kevin Leugers
Pathik Rupwate

Team Members

Linxi Wu

Data Scientist &
Project lead

Jiayi Du

Data Scientist

Pathik Rupwate

Data Engineer

Kevin Leugers

Data Engineer

Executive Summary

- **The Problem**
  - Many restaurants don't have massive budgets for rich data sets and analysis on competitive landscape and local factors
  - As a consulting firm, we have a limited hours and so are aiming to create a reusable solution that scales

- **The Solution**
  - Proof of concept tested for an example case in Austin, we help our client identify promising locations and impactful local factors to consider
  - We create a data pipeline and model that can be easily applied to other cities in the United States. We also create analyses and a dashboard that can be easily repurposed

Research Questions:

- **Questions:**
  - Which ZIP Codes are most attractive to open a restaurant of a specific type?
  - Will crime in the local area impact our rating?
  - For a given area of interest, what competition is in the area?
  - What specific restaurant types are preferred in a city? Which areas are they present/absent?

Agenda

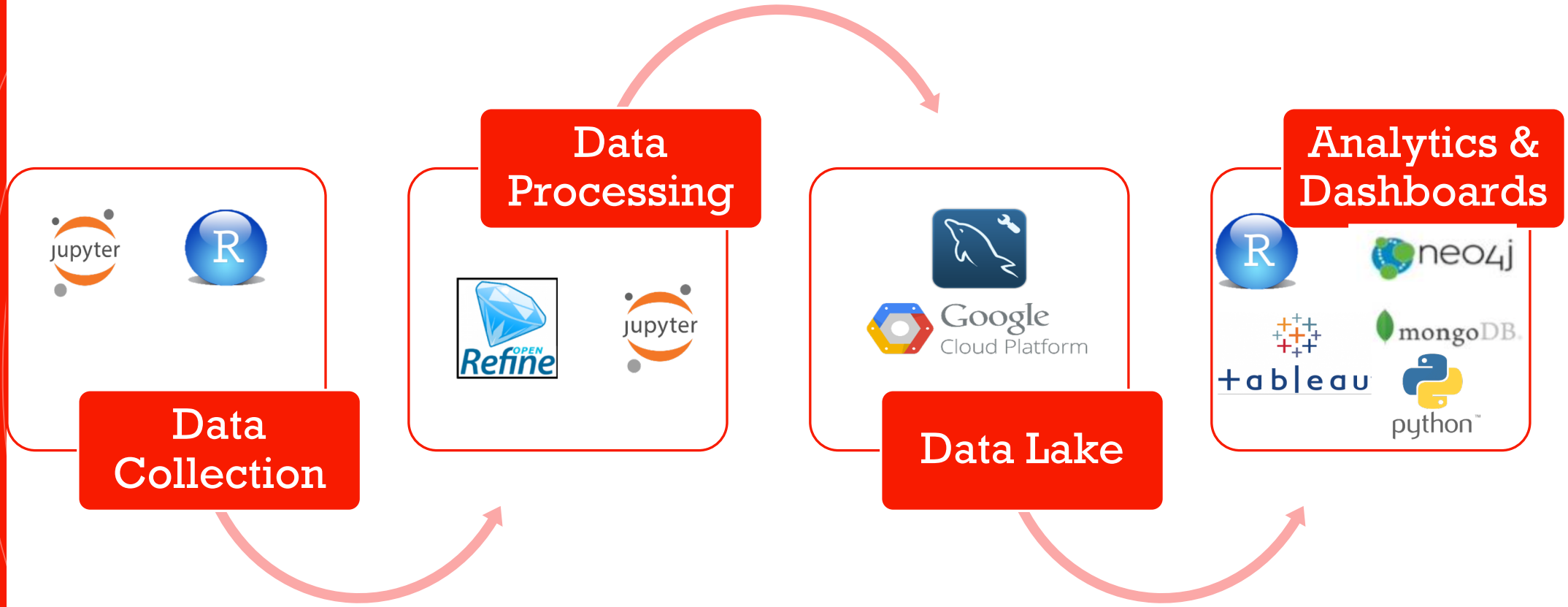
- Executive Summary
- Data Profile and Injection
- Data Cleaning
- Data Models
- Insights
- Recommendations

Data Profile

Data Profile

Data Source	Description	Size	Rows/Col
Austin Annual Crime Data	*Crime type by ZIP Code and date*	2.2MB	37462/5
Austin Neighborhoods by zip code	*Zip Codes and their associated neighborhoods in Austin*	5 KB	24/2
US Income	*Average income by ZIP Code and breakdown of population by income bracket*	475 KB	9738/5
Yelp business data	*Reviews, name, star rating, business category, lat/long, ZIP Code by business ID*	2.6GB	160,500/30
Unique ZIP Code List	*Exhaustive list of US Postal Codes*	350KB	33,120/1

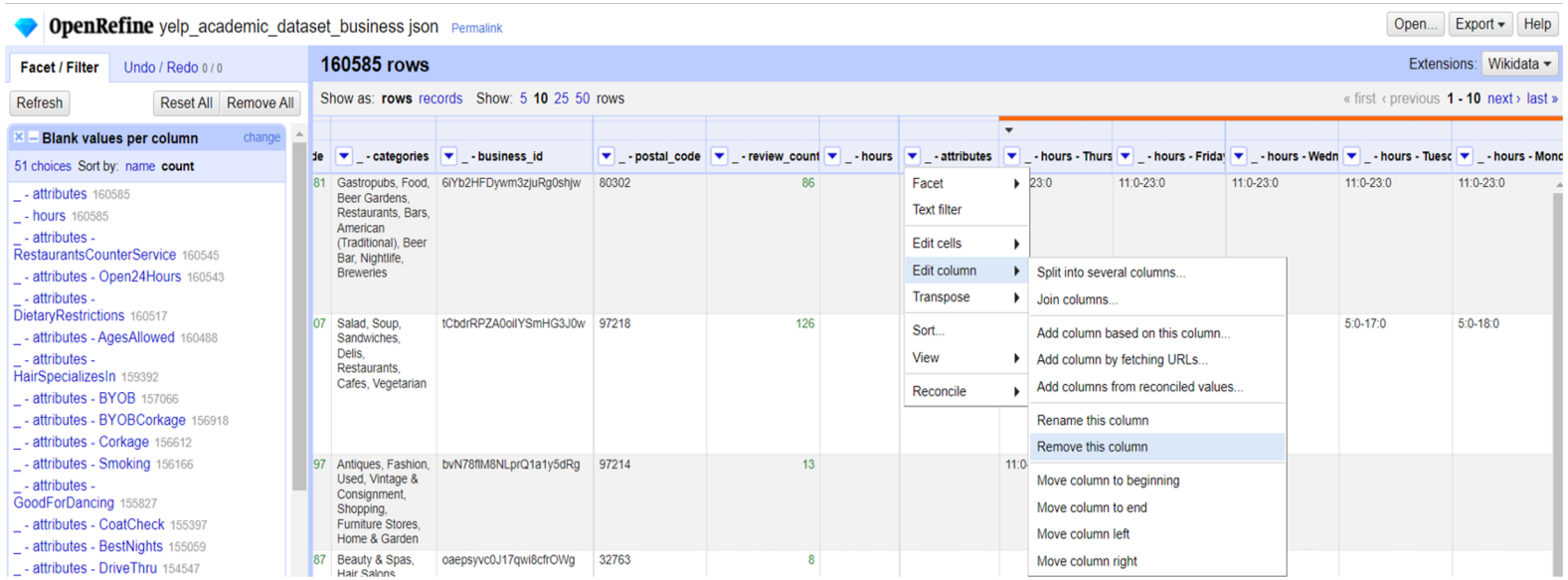
Data Implementation Tools



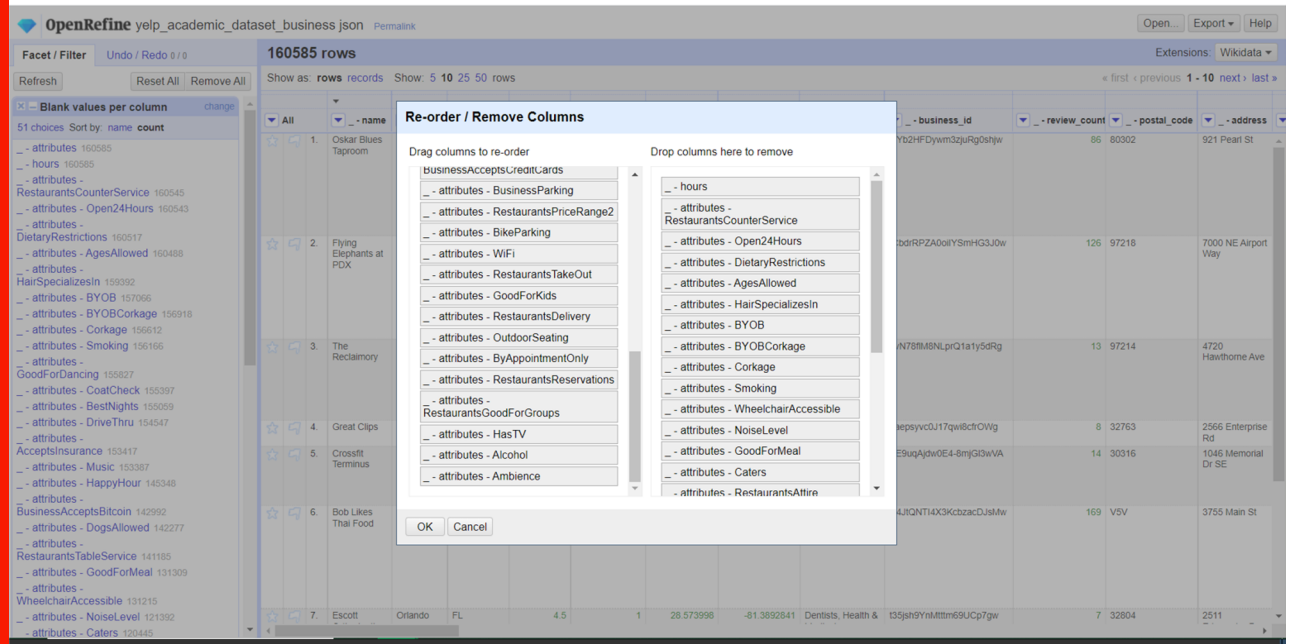
Data Ingestion

- We worked with very large data sets with over 160,000 rows. The large data files were cleaned with Openrefine due to its speed and ability to export files as csv, json, sql, and text.
- We also used sql and R to transform the tables into more understandable and usable tables.
- We cleaned empty cells, renamed or dropped columns, corrected spelling errors and narrowed down on Austin, Texas as our city of choice to analyze due the large sample size of Yelp data from Austin, Texas.

```
In [14]: !pip3 install beautifulsoup4

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
import re
import time
from datetime import datetime
import matplotlib.dates as mdates
import matplotlib.ticker as ticker
from urllib.request import urlopen
from bs4 import BeautifulSoup
import requests
```



Data Cleaning




```r
rest_austin <- business%>%
  filter(city == "Austin") %>%
  mutate(categories_new = strsplit(categories, split = ","))%>%
  unnest(cols = c(categories_new)) %>%
  filter(categories_new == "Restaurants")
```

```r
rest_final <- transform(rest_austin, zipcode = as.numeric(postal_code))
```

```
db.nbh.find({})
    .projection({})
    .sort({_id:-1})
    .limit(100);

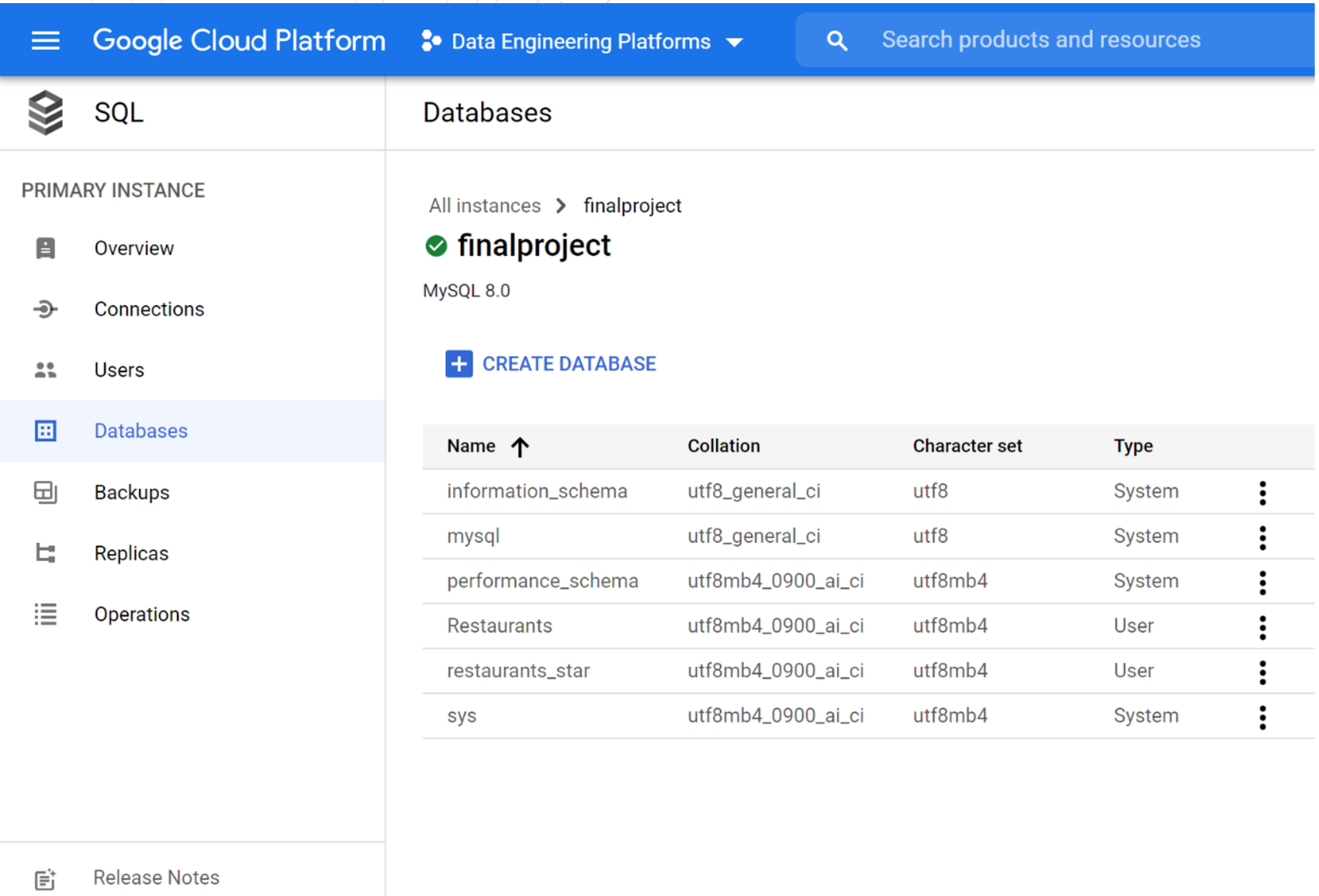
db.nbh.updateMany({},
     {$rename:{"ZIP Code":"postal_code"}});
```

nbh | 0.044 s | 27 Docs

	_id	postal_code	Neighborhood
1	61b3d055353e032c	78,757 (78.8K)	Central Austin
2	61b3d055353e032c	78,756 (78.8K)	Central Austin
3	61b3d055353e032c	78,751 (78.8K)	Central Austin
4	61b3d055353e032c	78,703 (78.7K)	Central Austin
5	61b3d055353e032c	78,705 (78.7K)	Central Austin
6	61b3d055353e032c	78,701 (78.7K)	Downtown
7	61b3d055353e032c	78,722 (78.7K)	East Austin
8	61b3d055353e032c	78,702 (78.7K)	East Austin
9	61b3d055353e032c	78,724 (78.7K)	Northeast Austin

Data Modeling & Database Design Considerations

Storage



GCP Selection Reasons:

- No on Prem Resources:
  - We don't have infrastructure and resources to maintain our own servers (among other limitations)
- Large Data Needs:
  - Yelp hosts an enormous amount of data.
- Expandable Storage Needs:
  - We anticipate expansion of our tools and data as adoption of these analyses grows so have chosen GCP.
  - Google Cloud Platform will be able to dynamically respond to our increased data needs

Database Design Considerations

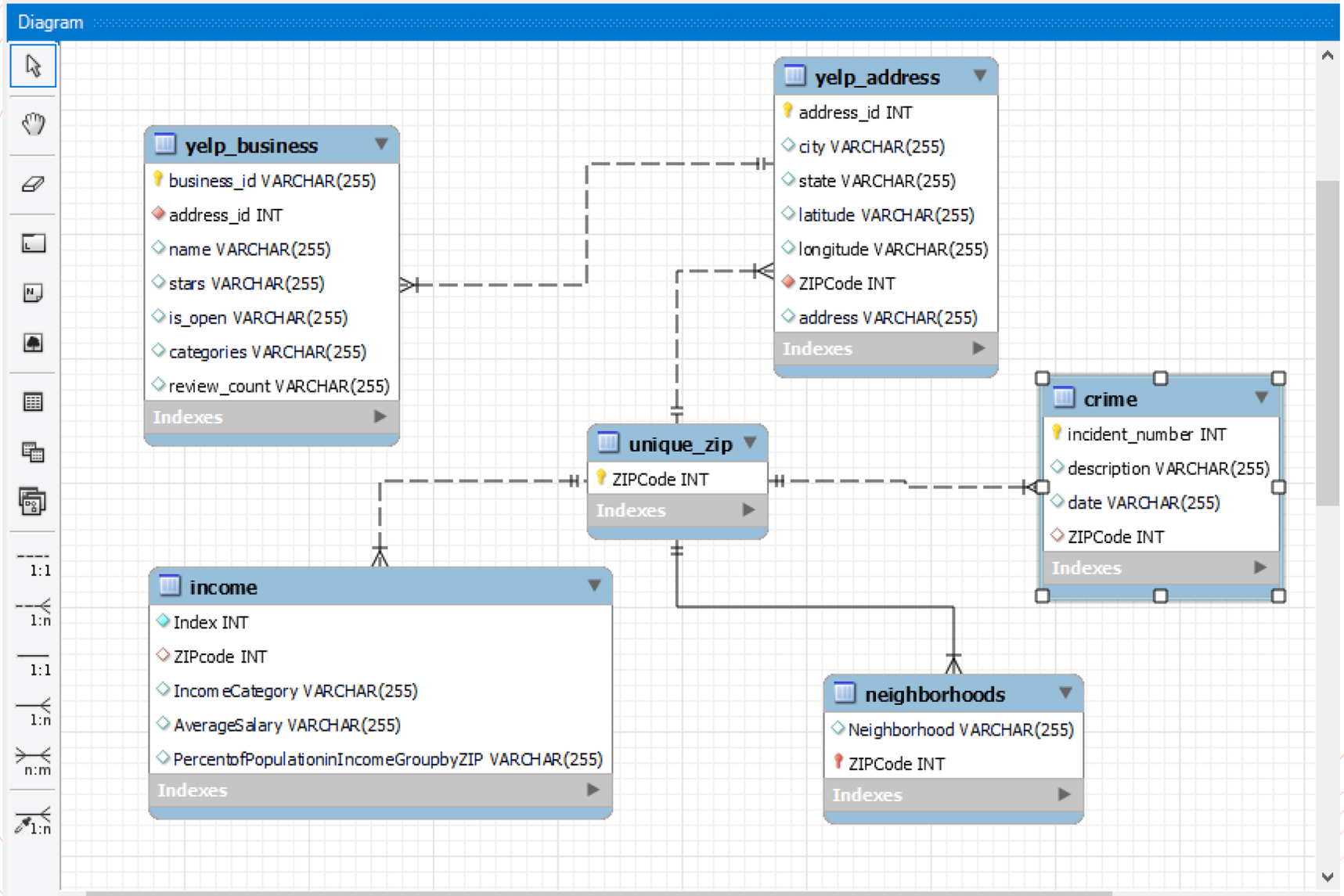
Model Selections:

- Transactional + OLAP
  - For efficient storage and future state preparation (where our client may ask to import first party data sets into our warehouse to manage), we have created a transactional model
  - However, the OLAP model will be the primary model leveraged for our reporting and analysis because the fact-dimension table set up lends itself to efficient querying

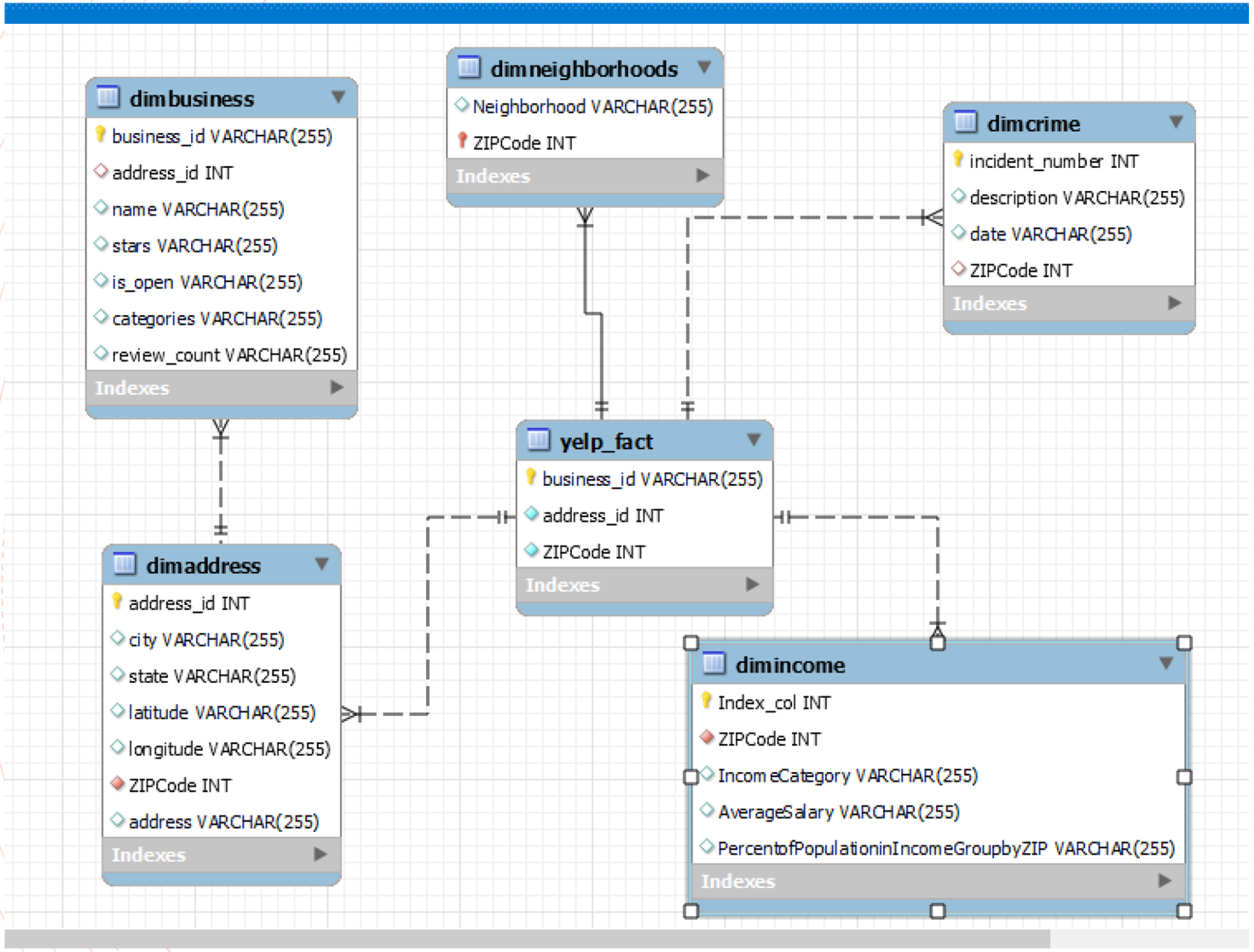
Model Creation:

- Transactional
  - 3NF: Most tables were in third normal form upon ingestion except for the Yelp data: we broke it out into a yelp_business table and yelp_address table to address dependent address fields
  - Foreign Key Connections: We brought in the unique list of US ZIP Codes to ensure the absence/presence of certain ZIP Codes in supporting tables would not exclude data or block foreign key constraints
- OLAP
  - Storage and Speed: For all data sets supporting the central Yelp data, we actively only bring in records with ZIP Codes in the Yelp data with dynamic subqueried SELECT statements during ingestion

Transactional Model



Dimensional Model



MongoDB

yelp

```
//Join neighbourhoods and restaurants on zip code.
//We see that only 1007 out of 1303 restaurants
//fall in Zip Codes in Austin texas
db.rest.aggregate([{
    $lookup:{
        from:'nbh',
        localField:"postal_code",
        foreignField:"postal_code",
        as: 'Neighborhood'
    }
},
{$unwind:"$Neighborhood"},
{$addFields:{Neighborhood:"$Neighborhood.Neighborhood"},
}]).sort({ "stars": -1 }).limit(10000);
```

rest | 0.332 s | 1,007 Docs | 50 | p. 1 | 1 - 50 | Table

	_id	categories	state	stars	review_co	postal_code	Neighborhood	name	longitude	latitude	is_open	city	business_id	address
16	61b3dfa8353e032c1c0(	Restauran	TX	5	7	78,722 (78.7K)	East Austin	DIY Thai Food	-97.7073	30.2965	0	Austin	ImiLLRt0wlWZXs7iP3vBiA	4209 Airpo
17	61b3dfa8353e032c1c0(	Restauran	TX	5	8	78,702 (78.7K)	East Austin	Mama Kong Cambodian Sc	-97.7164	30.2618	1	Austin	GQVxN1IJBzD2VFPxFGSx	2211 Webb
18	61b3dfa8353e032c1c0(	Restauran	TX	5	12	78,722 (78.7K)	East Austin	Luv Fats Ice Cream	-97.7073	30.2965	1	Austin	k6GZnr_QVIKTWsO4LNeM	4209 Airpo
19	61b3dfa8353e032c1c0(	Restauran	TX	5	14	78,702 (78.7K)	East Austin	Ararat ToGo + Coffee Wind	-97.7162	30.2592	1	Austin	iYGaz8ezrhxq5tXITMBcDQ	2401 East
20	61b3dfa8353e032c1c0(	Restauran	TX	5	5	78,701 (78.7K)	Downtown	Mike & Mike's Chicago Dog	-97.7425	30.2699	0	Austin	LtNCzkQDZ2ybUzHdzhGqI	722 Congr
21	61b3dfa8353e032c1c0(	Restauran	TX	5	89	78,757 (78.8K)	Central Austin	Hot Rod Coffee Trailer	-97.7391	30.3418	0	Austin	BkW7DbV3_1BPItPOTWe8	6546 Burn

```
//Number of Zip codes in Austin, Texas
db.rest.aggregate([
    {$set:{_id:"$postal_code"}},
    {$unionWith: {coll:"nbh", pipeline:[{$set:{_id:"$postal_code"}}]}},
    {$sort:{stars:1}},
    {$group:{_id:"$postal_code"}}
]).count();
```

0.047 s

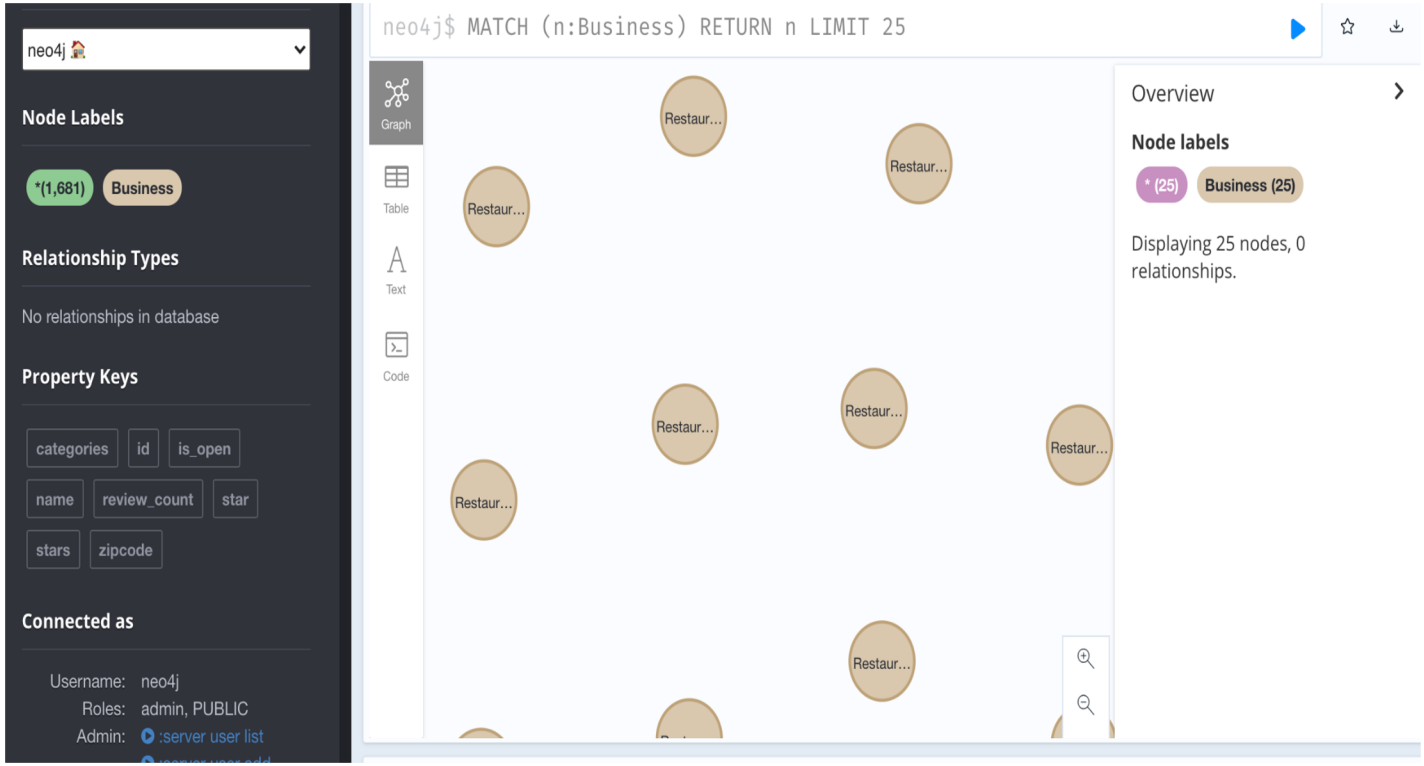
```
52
```

```
//Find all restaurants with a rating > 4 stars and sort by rating.
db.rest.aggregate([
    {$match: {stars:{$gt:4}}},
    {$project:{name:1, stars:1}},
    {$sort : {stars:-1}}
])
```

rest | 0.023 s | 283 Docs

	_id	stars	name
1	61b3dfa8353e032c1c0(	5	Little Sub Trailer
2	61b3dfa8353e032c1c0(	5	Celestial Pizza
3	61b3dfa8353e032c1c0(	5	SXSE Food
4	61b3dfa8353e032c1c0(	5	Hawai'i Nei cafe

Neo4j



1. Each node contain information of restaurant including star, name, zip code, review_count and is_open.
2. Each business node is in relationship with a crime node that contain crime number and a income node that contain the average income of the zip code area.

Visualization & Insights

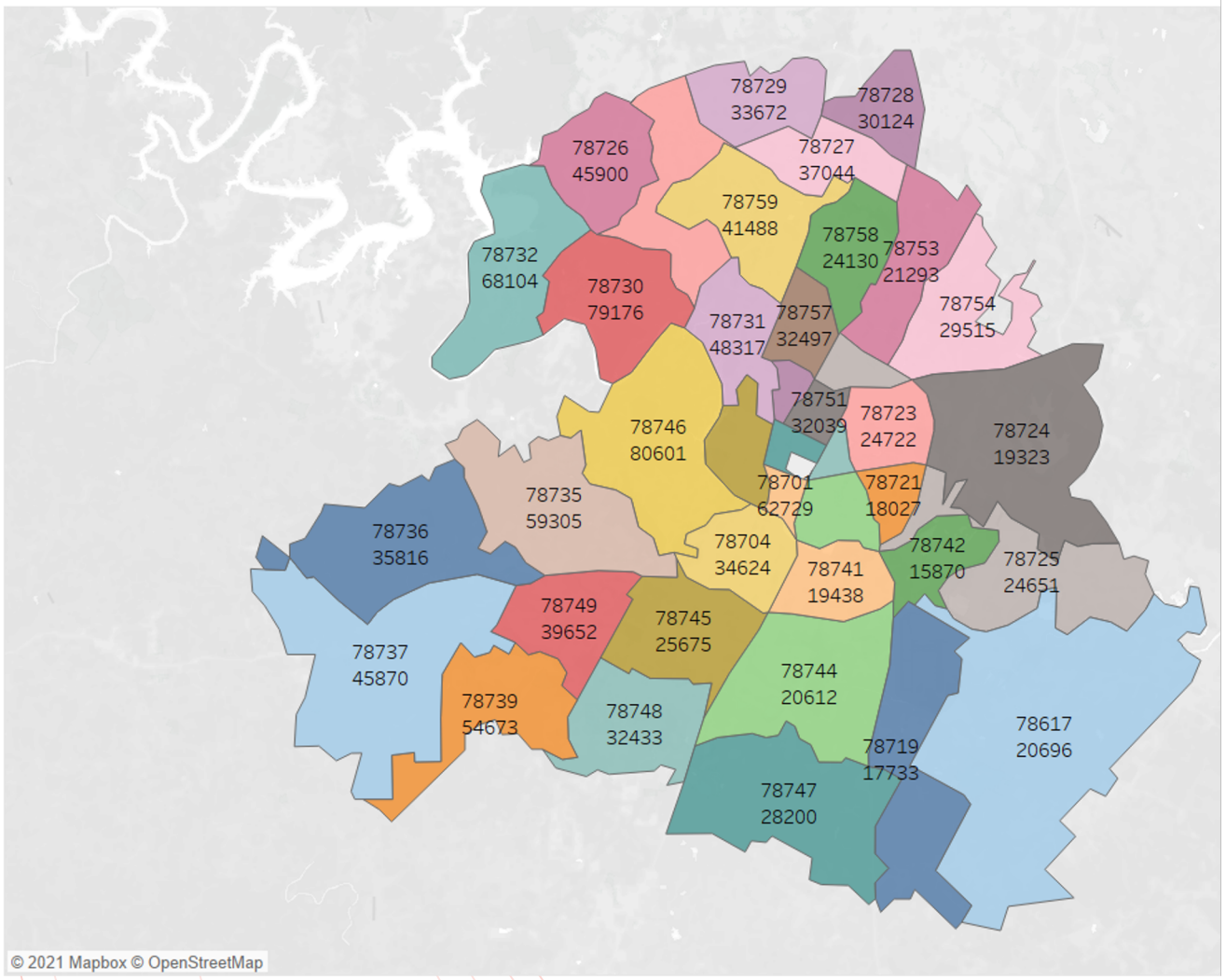
Statistical Analysis

Correlation analysis

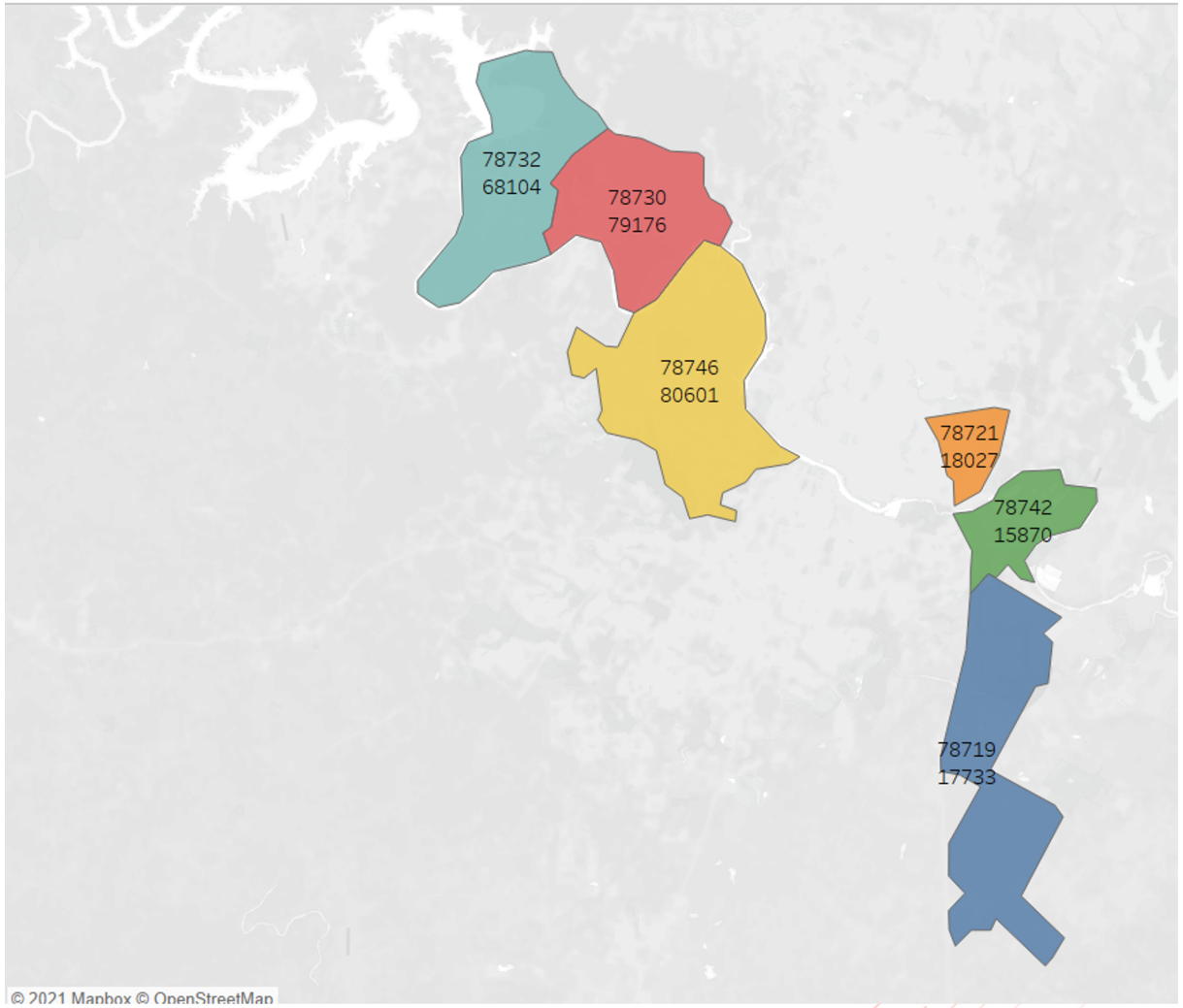
```
#Correlation analysis
```{r}
cor(correlation_data$stars, correlation_data$crime_num)
cor(correlation_data$stars, correlation_data$review_count)
cor(correlation_data$crime_num, correlation_data$salary)
```
```

```
[1] 0.1557746
[1] NA
[1] 0.1473701
```

Linear Regression

```
```{r}
model <- final %>%
  lm(stars ~ crime_num, data = .)
summary(model)
```
```

The regions with the highest and lowest income

Salary with Zipcode



Salary with Zipcode



The regions with the severe crime incidents

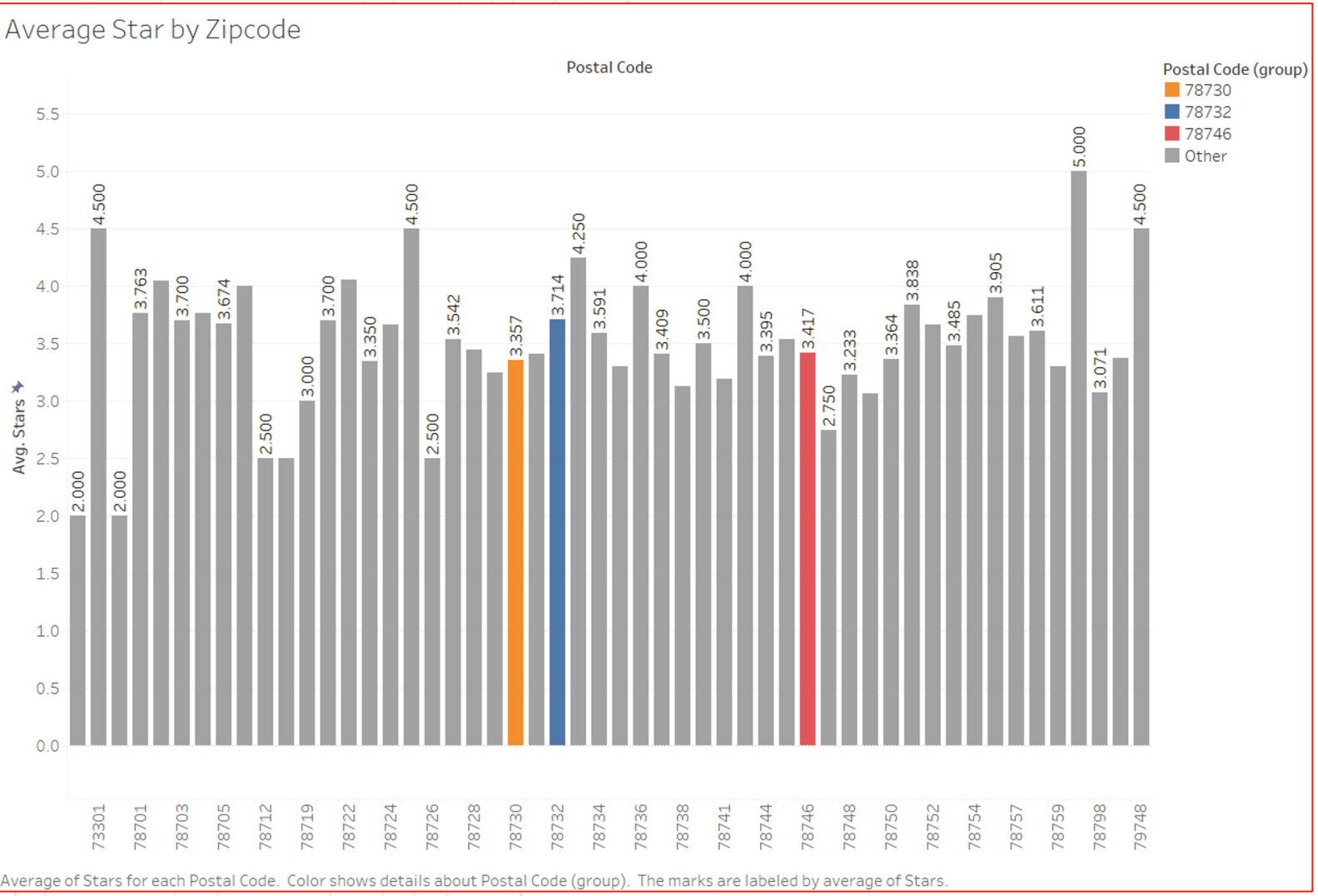
Assumptions:

- The severe crime such as murder or assault will impact the restaurant business in the region

Insights:

- The severe crime incidents such as murder or assault mostly happened in the dowtown area
- The three regions with highest income and the three with lowest income have relatively lower severe crime rate than downtown region

Average Star by Zip Code



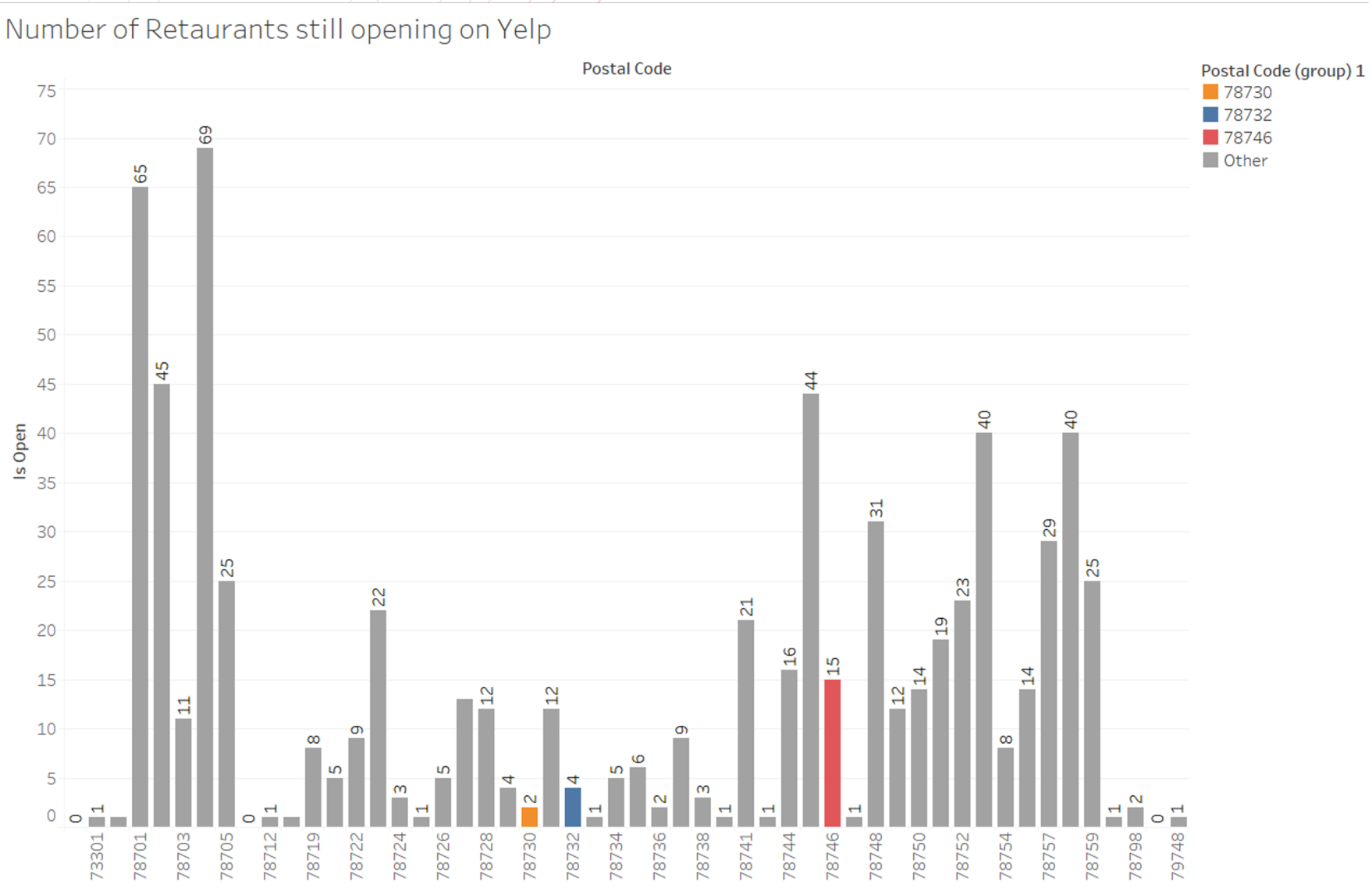
Business Logics:

- The new restaurant opened in the regions with highest Yelp rating might suffer fierce competitions
- Finding the region with a fairly averaged rating will benefit the restaurant owners

Insights:

- The places with the highest income level actually have a Yelp rating that fits the needs

Number of Restaurants still opening

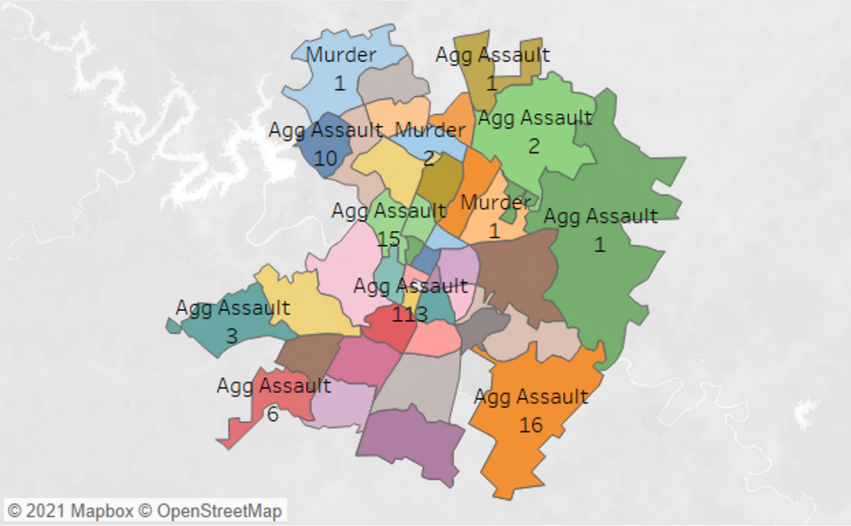


Insights:

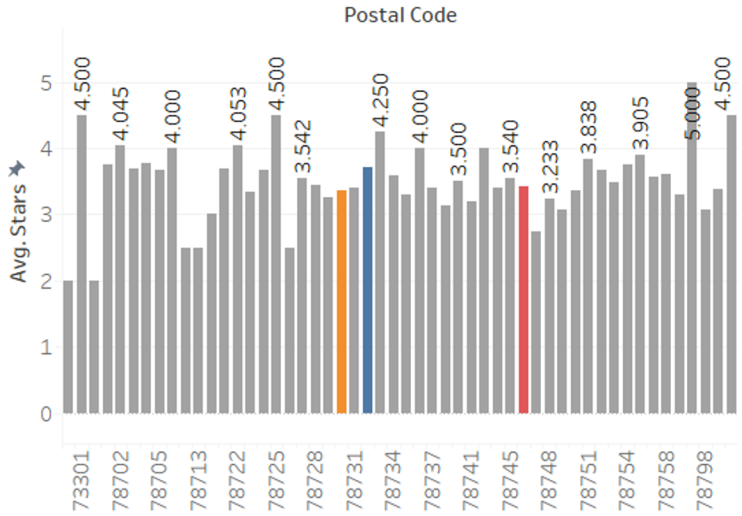
- Following the logic of last slide, we also found that the number of restaurants opening in those three regions with highest income is at the average level
- The downtown area has the highest number of restaurants opening,which indicates higher competition

A Tableau Dashboard to help keep trends

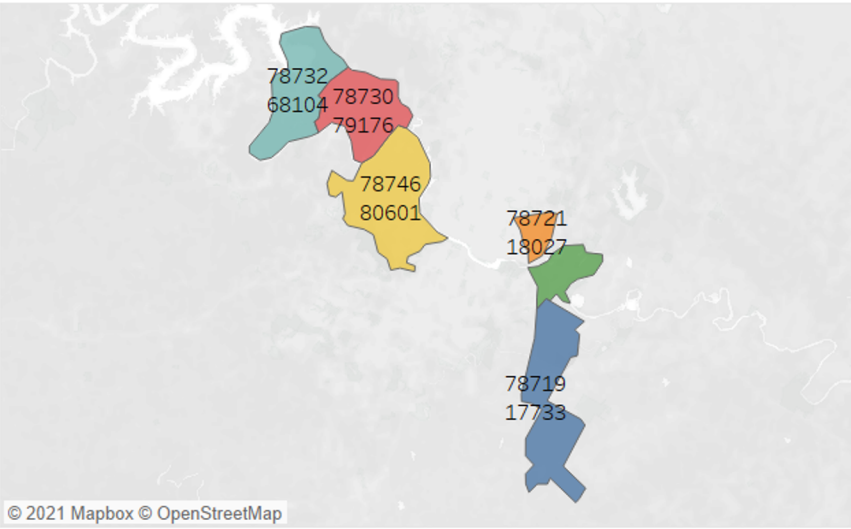
Crime with Number



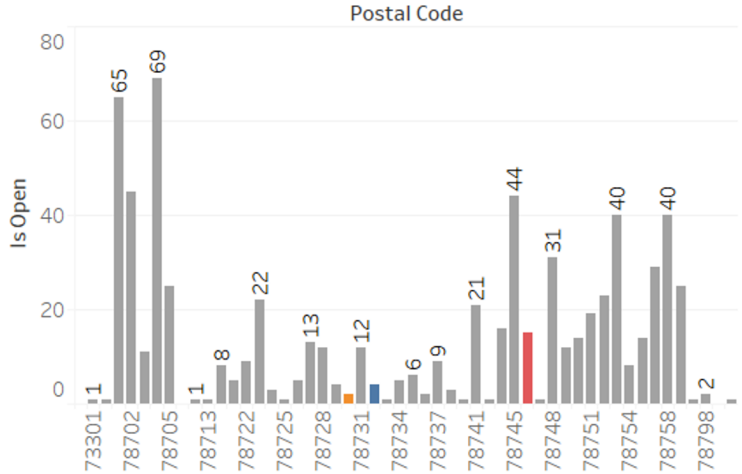
Average Star by Zipcode



Salary with Zipcode



Number of Retaurants still opening on Yelp



Benefits:

- As the covid-19 pandemic fading away, the trends in the restaurants might change. Therefore, having a Tableau Dashboard on hand is important to make a swift change.

Recommendations

Recommendations

- The restaurant owners should follow the insights from the crime data, average income by zip code, and the Yelp data to make decisions
- The optimal place to open a restaurant is in region with a zip code of 78746
- The region we recommend has the highest average income, has the relatively low crime incidents, and average level of completion

Thank You